Plant Syst Evol (2009) 282:71–86
DOI 10.1007/s00606-009-0208-2

ORIGINAL ARTICLE

What phylogeny and gene genealogy analyses reveal about homoplasy in citrus microsatellite alleles

Noelle A. Barkley · Robert R. Krueger · Claire T. Federici · Mikeal L. Roose

Received: 27 June 2008 / Accepted: 30 June 2009 / Published online: 28 July 2009

Abstract Sixty-five microsatellite alleles amplified from ancestral citrus accessions classified in three separate genera were evaluated for sequence polymorphism to establish the basis of inter- and intra-allelic genetic variation, evaluate the extent of size homoplasy, and determine an appropriate model (stepwise or infinite allele) for analysis of citrus microsatellite alleles. Sequences for each locus were aligned and subsequently used to determine relationships between alleles of different taxa via parsimony. Interallelic size variation at each SSR locus examined was due to changes in repeat copy number with one exception. Sequencing these alleles uncovered new distinct point mutations in the microsatellite region and the region flanking the microsatellite. Several of the point mutations were found to be genus, species, or allele specific, and some mutations were informative about the inferred evolutionary relationships among alleles. Overall, homoplasy was observed in alleles from all three loci, where the core microsatellite repeat was changed causing alleles of the same size class to be identical in state but not identical by descent. Because nearly all changes in allele size (with one exception) were due to expansion or contraction of the repeat motif, this suggests that a stepwise mutation model, which assumes homoplasy may occur, would be the most appropriate for analyzing *Citrus* SSR data. The collected data indicate that microsatellites can be a useful tool for evaluating *Citrus* species and two related genera since repeat motifs were reasonably well retained. However, this work also demonstrated that the number of microsatellite alleles is clearly an underestimate of the number of sequence variants present.

Keywords Homoplasy · *Citrus* · Microsatellite alleles · Parsimony · Gene network · Sequencing

N. A. Barkley
Graduate Program in Genetics, Genomics, and Bioinformatics, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

R. R. Krueger
USDA-ARS National Clonal Germplasm Repository for Citrus and Dates, Riverside, CA 92507, USA

C. T. Federici · M. L. Roose
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

N. A. Barkley (✉)
1109 Experiment Street, Griffin, GA 30223, USA
e-mail: elle.barkley@ars.usda.gov

Introduction

Microsatellites, also known as simple sequence repeats (SSRs), have been widely utilized in molecular genetic studies for mapping, fingerprinting, genetic diversity, and phylogenetic reconstruction. These markers are characterized by a 1- to 6-bp core repeat that is tandemly repeated in the genome and are generally thought to arise by DNA slippage during DNA synthesis (Schlötterer 1998). These tandem repeats are ubiquitous and can be found in nuclear, chloroplast, and mitochondrial genomes. One of the main properties of microsatellite markers that makes them so widely used in genetic research is that the polymorphism level can be highly discriminating, sufficient enough to display unique, specific genotypes for each individual in a population from relatively few markers (Estoup et al. 2002). This extreme polymorphism is a consequence of the high mutation rates of these sequences, which allow variability in species otherwise characterized by low levels of genetic diversity (Peakall et al. 1998).

The precise manner in which microsatellite loci mutate is not clear and can differ by species or loci evaluated. Two molecular mechanisms are thought to play a role in the rapid formation of new alleles at microsatellite loci: unequal exchange in meiosis and slipped-strand mispairing in replication (Levinson and Gutman 1987; Valdes et al. 1993; Orti et al. 1997; Zhu et al. 2000). These mutational mechanisms can generate allelic homoplasy whereby alleles are identical in state (or length), but not identical by descent, and thus contain different sequence motifs. The amount of allelic homoplasy in microsatellite loci, however, seems likely to depend on various factors such as time since divergence and mutation rate. Homoplasy in microsatellite alleles causes apparent similarity, but in reality, masks true evolutionary differences (Angers et al. 2000) among alleles. This phenomenon has often been characterized as the "noise," whereas homology can be characterized as the evolutionary "signal" (Estoup and Cornuet 1999). Apparent confusion between homology and homoplasy, which can easily occur in microsatellite studies based solely on allele size data, can potentially lead to inaccurate measures of genetic diversity, population divergence, relatedness, phylogenetic reconstruction, and inaccurate interpretation of population structure (Viard et al. 1998; Taylor et al. 1999). However, the exact effects that a given amount of homoplasy will have on the parameters used to describe population structure are not very clear and are difficult to predict (Rousset 1996; Orti et al. 1997).

Calculating genetic distance between individuals using microsatellite data depends on evolution models that aim to replicate the complex mutational process occurring at microsatellite loci (Buschiazzo and Gemmell 2006). Two molecular evolution models commonly used for the analysis of SSR markers are the stepwise mutation model (SMM) and the infinite allele model (IAM). The IAM postulates that each new mutation produces a unique allele (Kimura and Crow 1964), which allows for the creation of an infinite number of allelic states not already present in the population (Estoup and Cornuet 1999; Anmarkrud et al. 2008). On the other hand, the SMM assumes that there is equal probability of gaining or losing a single repeat unit within the microsatellite region to produce new distinguishable alleles. Unlike the IAM, SMM takes into account mutations back to a previous state (Anmarkrud et al. 2008). The SMM assumes that mutations result in alleles that have similar repeat units to the alleles from which they were derived. It further assumes that the differences in repeat units are informative in regard to the amount of time that has passed since the two alleles shared a common ancestor. Genetic distances based on the IAM, however, ignore this information (Goldstein et al. 1995).

Currently, the available data and simulations seem to suggest that most microsatellite sequences change in a stepwise manner (Valdes et al. 1993; Shriver et al. 1995; Zhu et al. 2000; Estoup et al. 2002). However, microsatellite alleles in maize do not always change in a stepwise manner. Most of the polymorphism in allele sizes detected was due to indels in the regions that flank the microsatellite repeat (Matsuoka et al. 2002). Twenty *Arabidopsis* microsatellite loci were evaluated and found to have complex mutational patterns that did not fit either the SMM or IAM consistently (Symonds and Lloyd 2003). Allelic size variation was also investigated in an intergeneric study of puma (*Puma concolor*) and the domesticated cat (*Felis catus*). This study showed that 80% of comigrating alleles between these two species displayed size homoplasy. The sequence differences between alleles of homologous puma and domestic cat microsatellite loci raised doubts about the accuracy of microsatellite-based phylogenetic comparisons between these distantly related mammalian genera (Culver et al. 2001). Overall, several exceptions to a strict SMM have been reported in which more complex mutations occurred among alleles, which altered the sequence content of alleles that were identical in state (Chen et al. 2002; Curtu et al. 2004; Hua et al. 2006; Lia et al. 2007).

Microsatellite markers have been the most commonly utilized markers in molecular biology for mapping, genetic diversity, phylogenetic construction, and fingerprinting because they are codominant, highly polymorphic, and easy to use. Frequently, these markers are developed from sequences containing repeat elements discovered in a particular species of interest, but employed across multiple related species or genera. Chen et al. (2002) demonstrated that allele size is an adequate measure of genetic difference when working with plants that are very closely related. However, when phylogenetic or evolutionary inferences are employed with distantly related species, then evaluation and verification of the SSR allele via sequencing is necessary because hidden motifs in alleles that are identical in state have been detected.

In 2006, the genetic diversity and phylogenetic relationships of multiple *Citrus* species and two related genera using a set of 24 microsatellite markers derived from *C. maxima* were assessed (Barkley et al. 2006). Therefore, the scope of this work was to verify the sequence content of citrus microsatellite alleles derived from 11 different species in three separate genera to examine the nature of variability among different-sized alleles, evaluate the extent of homoplasy among citrus SSR alleles in order to assess their utility for measuring phylogenetic relationships, and assess if the repeat motif is retained when using SSR markers over a broad range of taxonomically divergent *Citrus* species and related genera.

Materials and methods

Allele and taxa selection

Citrus and its close relatives are represented by 28 genera in the tribe Citreae of the subfamily Aurantioideae in the family Rutaceae (Swingle and Reece 1967). There are two commonly used classifications of *Citrus*: Swingle (Swingle and Reece 1967), and Tanaka (Tanaka 1977). Swingle lumps species together, recognizing 16 species in the genus *Citrus*, whereas Tanaka splits species, recognizing 162 *Citrus* species. The difficulty in classifying *Citrus* taxa is mainly due to repeated cross-pollination and to adventitious nucellar embryony, which stabilizes and perpetuates hybrid taxa (Scora 1975). Scora (1975) and Barrett and Rhodes (1976) suggested that there are only three "basic" true species of *Citrus* within the subgenus *Citrus* as defined by Swingle: citron (*C. medica* L.), mandarin (*C. reticulata* Blanco), and pummelo (*C. maxima* L. Osbeck). Nearly all *Citrus* species freely hybridize with one another, and thus, Mabberley (1997) suggests that taxonomic rank has been inflated due to the commercial importance of this crop and that only three species (*C. medica*, *C. maxima*, and *C. reticulata*) should be recognized for the subgenus *Citrus*. Other cultivated *Citrus* species within the subgenus *Citrus* are believed to be hybrids derived from these true species, species of the subgenus *Papeda*, or closely related genera, ideas generally supported by molecular marker data (Federici et al. 1998; Nicolosi et al. 2000; Barkley et al. 2006).

Given the taxonomy and prevailing theory on many *Citrus* species being derived by natural hybridization, alleles were carefully chosen from the data set of Barkley et al. (2006), which examined genetic diversity in a population of 370 citrus accessions and its relatives using SSR markers. Thirty-nine alleles were sampled from accessions considered to be "true" ancestral species within the subgenus *Citrus* [including *C. medica* (n = 9), *C. maxima* (n = 17), and *C. reticulata* (n = 13)], 5 alleles sampled from the subgenus *Papeda*, and 18 alleles sampled from their closest relatives, *Poncirus trifoliata* [trifoliates (n = 10)] and *Fortunella* spp. [kumquats (n = 8)] (Table 1). The remaining three samples were derived from hybrid taxa. The criterion for choosing taxa containing a particular allele was to maximize the number of different ancestral taxa (species) containing the allele of interest when possible. The citrus relatives were included to help evaluate if the repeat motif was conserved when crossing what is assumed to be more distant taxonomic borders. Additionally, alleles were selected that ranged from very low to high frequency (0.0108–0.8469) in the population (Barkley et al. 2006) to evaluate if allelic richness had any influence on intra-allelic variation (Table 2). In general, we did not sample known and probable hybrid taxa among naturally occurring forms since the goal of this study was to compare alleles derived from ancestral taxa. Thus, alleles chosen in this study were selected because they occurred frequently in putative ancestral taxa.

PCR, cloning, and sequencing of SSR alleles

The three loci used for this study were cAGG9, CCT01, and GT03, which had 6, 7, and 19 alleles, respectively, with polymorphic information content (PIC) values of 0.478, 0.247, and 0.834 in a study of 370 *Citrus*, *Poncirus*, and *Fortunella* accessions. PCR and gel electrophoresis conditions were performed as described previously (Barkley et al. 2006). All SSR alleles were cloned following the instructions in TOPO TA Cloning kit from Invitrogen (Carlsbad, CA). The ligation reaction consisted of 2 µl of PCR product, 0.5 µl of 1.2 M NaCl, and 0.5 µl of plasmid at a concentration of 10 ng/µl. The plasmid used was pCR 2.1-TOPO, which was provided by Invitrogen in the cloning kit, and contains Topoisomerase I from *Vaccinia* virus covalently bound to the vector that catalyzes the ligation of the PCR product into the vector. Chemically competent *E. coli* cells were transformed by adding the entire ligated product (3 µl) to TOP10 cells provided by Invitrogen. The cells were spread onto pre-warmed LB agar plates (1% tryptone, 0.5% yeast extract, 1% NaCl, 1.5% agar adjusted to pH 7.0) containing 40 mg/ml X-gal (scorable marker) and 50 µg/ml of kanamycin (selectable marker) and incubated overnight (12–16 h) at 37°C to allow colonies to develop.

Colonies were screened visually for a lack of color. Ten white colonies per plate were screened for the presence of an SSR allele by amplifying the plasmid with M13 primers included in the cloning kit. The PCR consisted of 5.4 µl of dH_20, PCR buffer (1×), magnesium chloride (1 mM), dNTPs (0.2 mM), M13F and M13R (7.5 ng/µl), and a scraping of cells from a single colony. The thermocycling conditions included a 2-min denaturing step at 92°C for 1 cycle; 30 cycles of 92°C for 30 s, 52°C for 30 s, and 72°C for 1 min; and a final elongation cycle of 72°C for 7 min. The PCR products were separated on a 4% precast agarose E-gel (Invitrogen; Carlsbad, CA) and scored visually for the presence of an insert. Two size standards were run on each E-gel to determine the insert sizes (pGEM Promega; Madison, WI and 100-bp marker, Invitrogen). Positive colonies were grown in 3 ml of liquid LB media (1% tryptone, 0.5% yeast extract, and 1% NaCl adjusted to pH 7.0) overnight, and the plasmids were isolated following the instructions from a Qiagen (Valencia, CA) mini-prep kit. Plasmids were sequenced

Table 1 A list of accessions used in this study along with their respective allele sizes that were cloned and sequenced from markers CCT01, cAGG9, and GT03

CRC no.	Cultivar name	Genus species	Taxonomic group	Marker	Allele size (gel score)	Allele size (sequence)	GenBank accession no.
3056	Unnamed	*C.* sp.	Papeda	CCT01	155	158	EU182531
3793	Unnamed	*C.* sp.	Papeda	CCT01	155	158	EU182533
3797	Unnamed	*C.* sp.	Papeda	CCT01	155	158	EU182534
3780	Unnamed	*C. halimii*	Citron hybrid	CCT01	155	158	EU182532
0661	'Indian'	*C. medica*	Citron	CCT01	158	161	EU182525
3055	'Bengal'	*C. medica*	Citron	CCT01	158	161	EU182523
3237	Unnamed	*F. japonica*	Kumquat	CCT01	158	161	EU182526
0300	'Parson's Special'	*C. reticulata*	Mandarin	CCT01	158	161	EU182527
3816	'Kinkoji Unshiu'	*C. reticulata*	Mandarin	CCT01	158	161	EU182529
0644	'Philippine'	*C. maxima*	Pummelo	CCT01	158	161	EU182524
2355	'Kao Panne'	*C. maxima*	Pummelo	CCT01	158	161	EU182528
3066	'Sour'	*C. maxima*	Pummelo	CCT01	158	161	EU182522
2554	'Barnes'	*P. trifoliata*	Trifoliate	CCT01	158	161	EU182530
3055	'Bengal'	*C. medica*	Citron	CCT01	161	164	EU182519
3237	Unnamed	*F. japonica*	Kumquat	CCT01	161	164	EU182517
3816	'Kinkoji Unshiu'	*C. reticulata*	Mandarin	CCT01	161	164	EU182520
3793	Unnamed	*C.* sp.	Papeda	CCT01	161	164	EU182516
0448	'Moanalua'	*C. maxima*	Pummelo	CCT01	161	164	EU182518
0644	'Philippine'	*C. maxima*	Pummelo	CCT01	161	164	EU182521
2341	'Karn Lau Yau'	*C. maxima*	Pummelo	CCT01	161	164	EU182515
0578	'Fleming's Shaddock'	*C. maxima*	Pummelo	CCT01	164	167	EU182511
1225	'Hunnan'	*C. maxima*	Pummelo	CCT01	164	167	EU182510
2341	'Karn Lau Yau'	*C. maxima*	Pummelo	CCT01	164	167	EU182512
2355	'Kao Panne'	*C. maxima*	Pummelo	CCT01	164	167	EU183513
3066	'Sour'	*C. maxima*	Pummelo	CCT01	164	167	EU182514
0138	'Indian'	*C. medica*	Citron	cAGG9	103	105	EU182553
2875	'Japansche'	*C. medica*	Citron	cAGG9	103	105	EU182554
3527	'Hiawassie'	*C. medica*	Citron	cAGG9	103	105	EU182555
3163	'Indian wild orange'	*C. reticulata*	Mandarin	cAGG9	103	105	EU182551
0131	'Santa Barbara'	*C. limonia*	Rangpur	cAGG9	103	105	EU182552
3147	Unnamed	*C. reticulata*	Mandarin	cAGG9	112	114	EU182546
3150	Unnamed	*C. reticulata*	Mandarin	cAGG9	112	114	EU182548
3845	'King'	*C. reticulata*	Mandarin	cAGG9	112	114	EU182550
3469	'Hanayu'	*C. hanaju*	Papeda	cAGG9	112	114	EU182547
0131	'Santa Barbara'	*C. limonia*	Rangpur	cAGG9	112	114	EU182549
3055	'Bengal'	*C. medica*	Citron	cAGG9	115	117	EU182541
0279	'Clementine'	*C. reticulata*	Mandarin	cAGG9	115	117	EU182542
3816	'Kinkoji Unshiu'	*C. reticulata*	Mandarin	cAGG9	115	117	EU182544
3845	'King'	*C. reticulata*	Mandarin	cAGG9	115	117	EU182540
1208	'Roeding's Pink'	*C. maxima*	Pummelo	cAGG9	115	117	EU182543
3326	'Scarlet Emperor'	*C. reticulata*	Mandarin	cAGG9	115	117	EU182545
3816	'Kinkoji Unshiu'	*C. reticulata*	Mandarin	cAGG9	118	120	EU182538
1208	'Roeding's Pink'	*C. maxima*	Pummelo	cAGG9	118	120	EU182539
2240	'Siamese Acidless'	*C. maxima*	Pummelo	cAGG9	118	120	EU182537
2554	'Barnes'	*P. trifoliata*	Trifoliate	cAGG9	118	120	EU182535
4008	'Seedling'	*P. trifoliata*	Trifoliate	cAGG9	118	120	EU182536
2875	'Japansche'	*C. medica*	Citron	GT03	151	151	EU182574

Table 1 continued

CRC no.	Cultivar name	Genus species	Taxonomic group	Marker	Allele size (gel score)	Allele size (sequence)	GenBank accession no.
2867	'Calashu'	*C. reticulata*	Mandarin	GT03	153	153	EU182571
3789	Unnamed	*F. hindsii*	Kumquat	GT03	153	153	EU182573
3790	'BB 394'	*F. hindsii*	Kumquat	GT03	153	153	EU182572
1471	'Meiwa'	*F. crassifolia*	Kumquat	GT03	167	167	EU182566
3789	Unnamed	*F. hindsii*	Kumquat	GT03	167	167	EU182568
3818	'Meiwa'	*F. crassifolia*	Kumquat	GT03	167	167	EU182569
3833	'Meiwa'	*F. crassifolia*	Kumquat	GT03	167	167	EU182567
1224	Unnamed	*C. maxima*	Pummelo	GT03	167	167	EU182570
3878	'S-1'	*C. medica*	Citron	GT03	171	171	EU182562
0644	'Philippine'	*C. maxima*	Pummelo	GT03	171	171	EU182561
3947	'Suisho Buntan'	*C. maxima*	Pummelo	GT03	171	171	EU182563
3151	'Australian	*P. trifoliata*	Trifoliate	GT03	171	171	EU182564
3888	Unnamed	*P. trifoliata*	Trifoliate	GT03	171	171	EU182565
4006	'Seedling'	*P. trifoliata*	Trifoliate	GT03	173	173	EU182556
2554	'Barnes'	*P. trifoliata*	Trifoliate	GT03	173	173	EU182557
3351	'Fairhope'	*P. trifoliata*	Trifoliate	GT03	173	173	EU182560
3549	'Simmons'	*P. trifoliata*	Trifoliate	GT03	173	173	EU182559
3876	'English Dwarf'	*P. trifoliata*	Trifoliate	GT03	173	173	EU182558

Citrons, mandarins, pummelos, and papedas are thought to be ancestral species, whereas kumquats (*Fortunella*) and trifoliates (*Poncirus*) are classified as *Citrus* relatives

Table 2 Number of single-site polymorphisms observed at each allele size class

Marker	Allele size (bp)	Allele frequency	No. of single site polymorphisms	No. of taxa cloned/allele
CCT01	164	0.0108	2	5
CCT01	161	0.1206	7	7
CCT01	158	0.8469	10	9
CCT01	155	0.0108	3	4
CCT01 TOTAL	–	–	22	25
cAGG9	118	0.0380	0	5
cAGG9	115	0.6658	3	6
cAGG9	112	0.0353	1	5
cAGG9	103	0.1821	3	5
cAGG9 TOTAL	–	–	7	21
GT03	173	0.0309	1	5
GT03	171	0.2739	2	5
GT03	167	0.0225	2	5
GT03	153	0.0281	3	3
GT03	151	0.1236	–	1
GT03 TOTAL	–	–	8	19
GRAND TOTAL	–	–	37	65

Allele frequency was calculated from the data set of Barkley et al. (2006), which examined a population of 370 *Citrus* and related taxa with 24 SSR markers

bidirectionally at the University of California, Riverside, Genomics Institute Core Instrumentation Facility using an ABI 3100 DNA sequencer (16-capillary). Multiple clones of a single allele were sequenced for over 38% of the selected alleles in this study to evaluate PCR and sequencing errors.

Sequence alignments and tree construction

AlignIR 2.0 (LI-COR; Lincoln, NE) was used to trim out the vector sequence and construct a consensus sequence of the forward and reverse sequence reads. No discrepancies between bidirectional sequences were noted in the insert region. All sequences were aligned using ClustalX (Thompson et al. 1997). Alignments were performed with low and high gap penalties. The setting used for the pairwise alignment parameter was 10.00 for the gap opening and 0.10 for the gap extension penalty. The multiple alignment parameters were set to have a gap opening of 10.00 and a gap extension penalty of 0.20. The pairwise alignment parameter was repeated using 100 for the gap opening and 7.5 for the gap extension penalties. The multiple alignment parameters were increased to 100 for the gap opening and 3.0 for a gap extension penalty. Changing the gap penalty parameters did not affect the sequence alignment. In the microsatellite region, the ClustalX alignment was also visually inspected and manually edited to minimize the number of gap locations.

Unweighted parsimony analysis with PAUP version 4.0 beta 10 was used to construct phylogenetic trees for each SSR marker (Swofford 2003). Parsimony searches all possible trees and evaluates each tree for the minimum number of mutations. This analysis performs well when convergence is rare, sampling is dense, and individual branches are short (Holder and Lewis 2003). Bootstrapping analysis that tests clade stability by resampling the data with replacements was conducted with 10,000 replicates. All gaps in the sequence were treated as a fifth base as opposed to being treated as missing. The sequence data were edited to reflect a 1-bp change for each respective trinucleotide or dinucleotide gap that occurred. This change ensures that a di- or trinucleotide gap in the repeat element is not treated as multiple characters/events, which would over-inflate the number of informative characters in each sequence. Additionally, gene genealogies (networks) were constructed from the sequence data (nexus format) of all the alleles at each marker by utilizing TCS version 1.13 (Clement et al. 2000). All gaps were treated as a fifth base. Sequence alignments were imported into DnaSP (DNA sequence polymorphism) version 3.5 (Rozas and Rozas 1999) to calculate statistics on DNA sequence variation such as π and θ for the three SSR markers studied.

Results

A total of 65 alleles mainly derived from ancestral *Citrus* species and two closely related genera (*Poncirus* and *Fortunella*) were cloned and sequenced from three SSR markers (Table 1). The relatives of *Citrus* were included to examine how often the microsatellite is conserved when crossing distant taxonomic borders. Even though *Poncirus* and *Fortunella* species can be hybridized with the genus *Citrus*, there is little evidence to suggest a long history of natural gene exchange between the genera *Poncirus* and *Citrus*. Furthermore, previous phylogenetic data based on molecular markers demonstrate that *Poncirus*, *Fortunella*, and *Citrus* are divergent (Nicolosi et al. 2000; Barkley et al. 2006; Pang et al. 2007), although cpDNA sequences place *C. medica* as more distant from other *Citrus* spp. than these related genera (Bayer et al. 2009). The alleles sequenced in this study were amplified using primers that targeted two trinucleotide repeat loci, one compound and one imperfect (cAGG9 and CCT01), and one imperfect dinucleotide repeat locus (GT03). The collected sequence data for each marker were used to create dendrograms to evaluate inter- and intra-allelic relationships via parsimony. Dendrograms were constructed utilizing the entire sequence (microsatellite region and flanking region) and the flanking region alone to evaluate if the repeat motif or the flanking region contributed in determining evolutionary relationships among alleles.

Locus CCT01

A total of 25 microsatellite alleles were cloned and sequenced from the trinucleotide locus CCT01 (Table 1). The indels observed in these sequences consisted of 3-bp repeats and occurred only within the microsatellite region. Size homoplasy in which alleles are identical in state but not identical by descent was detected in the alleles sequenced. For example, the 158-bp allele from CRC 644 pummelo had a compound repeat motif of $(TCC)_3(ACC)_2(TCC)_2$ while the remaining 158-bp alleles had a slightly different imperfect repeat motif consisting of $(TCC)_3ACC(TCC)_3$. The 158-bp allele from CRC 644 pummelo could have either arisen from another 158-bp allele by a T-to-A point mutation or from a 164-bp allele by deletions of TCC from both sides of the ACC interrupt. Most of the 161-bp alleles had a repeat motif of $(TCC)_3(ACC)(TCC)_4$ followed by TCT, but one had $(TCC)_3(ACC)(TCC)_3(TCT)_2$, which may have arisen by a C-to-T mutation, or by loss of a TCC and gain of a TCT, creating a new repeat motif. The numerous single-site mutations at this locus were the basis of several cases of apparent homoplasy.

The sequence data were employed to generate a gene tree that recognized eight characters that were parsimony informative and 17 variable characters that were uninformative. The tree (Fig. 1a) that resulted was not well resolved. The alleles in this tree did not segregate into several clusters of alleles of the same size class as would be expected for a microsatellite locus in which variation was

due solely to change in repeat length. For example, the main polytomy of this tree included taxa with 155-, 158-, 161-, and 164-bp alleles, which clustered together even though there were differing numbers of trinucleotide gaps in these sequences. It is possible that a 9-bp gap between the smallest (155 bp) and the largest (164 bp) alleles, which was treated by PAUP as three mutational steps to reflect the trinucleotide repeat and very few informative characters, was not an adequate difference to sufficiently separate these four alleles.

Locus cAGG9

Twenty-one microsatellite alleles amplified from locus cAGG9 were cloned and sequenced (Table 1). The inter-allelic size variation observed at this locus was due to indels within the microsatellite repeat. Very few cases of apparent homoplasy were observed at this locus. A tree was constructed from the sequence data obtained from this marker (Fig. 1b). Only six informative characters were identified, indicating very little sequence divergence, which could be a result of the high degree of cross hybridization and stabilization of hybrids via nucellar embryony among *Citrus*. However, since hybridization and nucellar embryony would apply equally to all loci examined and this limited sequence divergence was not observed in other loci, this may suggest that this locus is affected by stabilizing selection. Most of the sequence divergence detected appeared as trinucleotide gaps in the microsatellite region as indicated by the decrease in the number of parsimony informative characters (from six to one) when the microsatellite region was removed from the data set (data not shown). All of the alleles of the same size class clustered together and were unresolved, indicating that these allele sizes specify evolutionary relationships at this marker. All of the 118-bp alleles clustered together and could not be resolved. These alleles were derived from two trifoliate orange accessions (*P. trifoliata*), ancestral *Citrus* taxa including a mandarin (*C. reticulata*), and two pummelos (*C. maxima*), which are divergent based on taxonomic classification and phylogenies produced from molecular marker data (Federici et al. 1998; Nicolosi et al. 2000, Barkley et al. 2006). The sequence for this allele was completely conserved with no single-site polymorphism observed among these divergent taxa. This sequence conservation suggests that this allele may be ancestral, and thus, was present before the genera *Poncirus* and *Citrus* separated. Another, less likely possibility is that these distantly related taxa evolved the same derived characters independently.

The taxa with the 115-bp allele were chosen for this study because they were ancestral and classified in three separate species in the genus *Citrus*, and therefore, are taxonomically divergent. Accessions derived from these ancestral species separate into distinct clades in previous studies using SSR, RFLP, AFLP, or RAPD markers (Federici et al. 1998; Nicolosi et al. 2000; Barkley et al. 2006; Pang et al. 2007). The 115-bp sequences were fairly conserved with only three single-site polymorphisms observed within this allele class (Table 2). The next group, consisting of the five 112-bp alleles, was also unresolved in this tree and supported with a bootstrap value of 64%. Four of the five taxa with a 112-bp allele had no detectable polymorphisms when compared to one another. The last main group was the 103-bp alleles. All of the 103-bp alleles shared a transversion in the flanking region compared to the remaining allele size classes. The branch supporting these alleles was highly supported with a bootstrap value of 94% (Fig. 1b).

The microsatellite region was removed from all alleles produced at this locus to examine the effect of the polymorphisms in the flanking region and to determine how they influence the resolution of the tree. The resulting tree was much less resolved, and most of the alleles (112, 115, and 118) could not be distinguished from one another (data not shown). Since the resulting tree was less resolved than the tree with the entire sequence, this suggests that the allele sizes at this marker do indicate evolutionary relationships between sequences when the entire sequence is employed. However, in four of the five 103-bp alleles, the polymorphisms in the flanking region played a role in inferring the evolutionary relationships, since this allele segregated from the others examined due to a transversion observed in the flanking sequence. Moreover, this analysis demonstrates that contrary to other studies such as Rossetto et al. (2002), utilization of the flanking sequence for this locus would not be an effective strategy to deduce evolutionary relationships in citrus taxa.

Locus GT03

Nineteen alleles from locus GT03 were cloned and sequenced (Table 1). No indels were found in the regions flanking the microsatellite. However, there were several point mutations in these sequences both in the microsatellite region and the flanking sequence that produced multiple cases of homoplasy. Several of the point mutations detected in the alleles of this locus were genus/species specific or specific to a particular allele class. A dendrogram was constructed from the sequence data at this marker (Fig. 1c). This marker displayed the highest number of parsimony-informative characters with 18 informative characters. Many alleles of the same size clustered together and could not be distinguished. There were four main groupings in this tree, each consisting of a different allele size class (173, 171, 167, 153).

All of the 173-bp alleles, which occurred in the *Citrus* relative *Poncirus trifoliata*, were unresolved. The 171-bp alleles split into two groups. One cluster contained the 171-bp alleles produced from accessions classified in the genus *Poncirus* while the other cluster contained 171-bp alleles produced from accessions classified in the genus *Citrus*. This split would be expected based on taxonomic classification and phylogenies based on molecular marker data. The two 171-bp alleles from *Poncirus* accessions shared a C-to-T transition that was also observed in all of the 173-bp *Poncirus* alleles. This transition was genus/species specific, occurring only in *P. trifoliata*. As a group, the trifoliate oranges all tend to be similar to one another (Fang et al. 1997) and are divergent from the genus *Citrus* (Nicolosi et al. 2000; Barkley et al. 2006; Pang et al. 2007). This may explain why these alleles of different sizes share the same point mutation in the flanking sequence. Additionally, the 171-bp alleles displayed homoplasy because the 171-bp alleles produced from *P. trifoliata* accessions have an imperfect $(GT)_3TTCT(GT)_{14}$ repeat motif, whereas the 171-bp alleles derived from the genus *Citrus* have a compound $(GT)_3TT(CT)_2(GT)_{13}$ repeat motif.

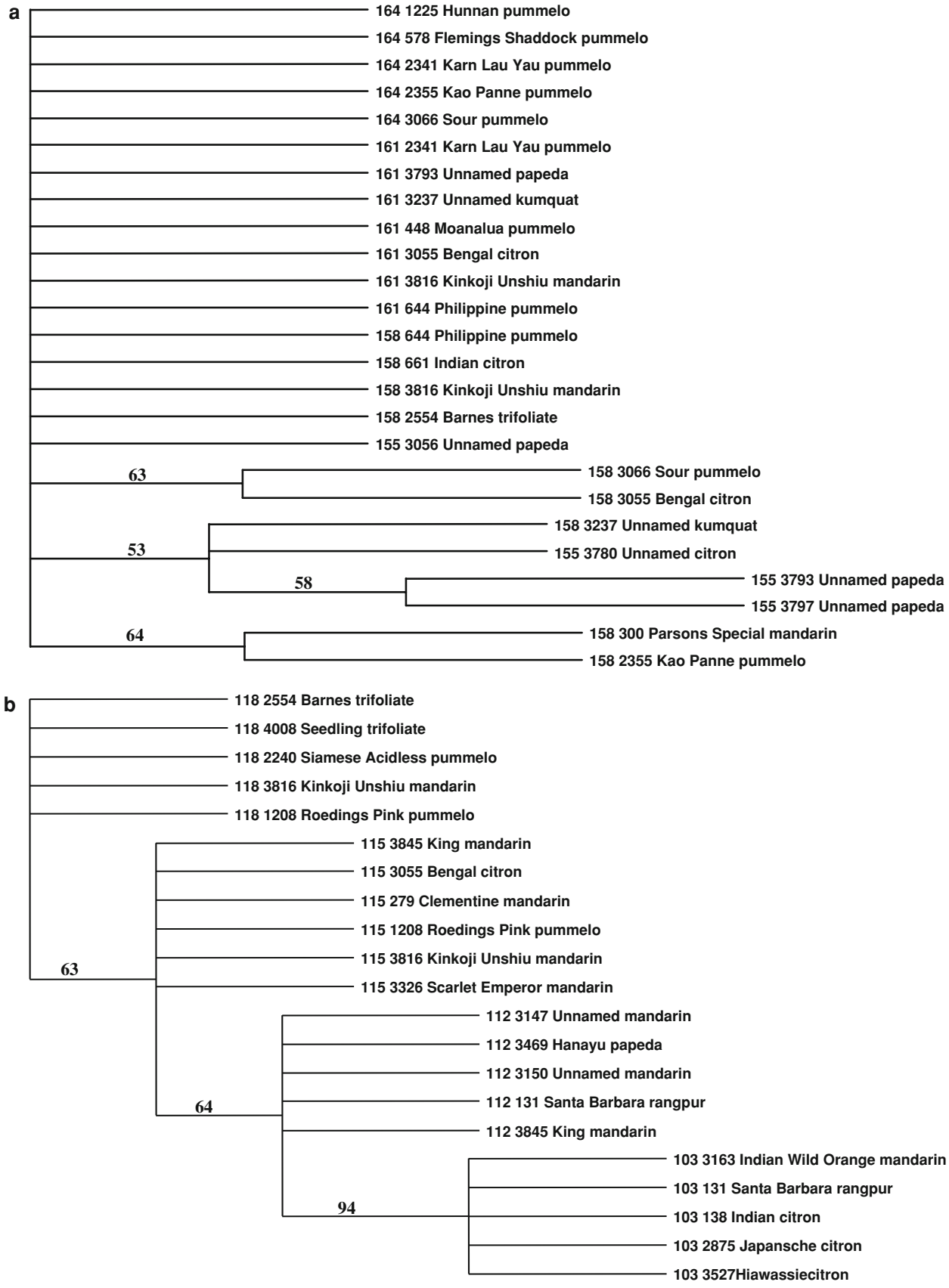
The 167-bp alleles were derived from four kumquats (two *Fortunella hindsii* and two *F. crassifolia*) and one pummelo (*Citrus maxima*). Transition mutations were detected within the microsatellite and flanking regions, respectively, that were specific to the genus *Fortunella* (kumquat), but not observed in the 167-bp allele from a pummelo (*C. maxima*). This produced two different microsatellite repeat motifs for the 167-bp alleles: $(GT)_3TTCT(GT)_{12}$ in pummelo and $(GT)_3TTCT(GT)_3AT(GT)_8$ in kumquats. The four kumquat accessions had no detectable sequence divergence; therefore these alleles clustered together with a bootstrap value of 68% and were completely unresolved (Fig. 1c). This divergence caused the pummelo accession with a 167-bp allele to cluster separately from the other 167-bp alleles. However, pummelo (*C. maxima*) accessions are classified in a different genus than the kumquats (*Fortunella*); therefore one might expect this based on the taxonomy. Additionally, *C. maxima* and *Fortunella* spp. have typically clustered in separate clades in previous molecular marker studies (Federici et al. 1998; Nicolosi et al. 2000; Barkley et al. 2006).

The last main group on this dendrogram consisted of three 153-bp alleles and one 151-bp allele produced from citron (*C. medica*), mandarin (*C. reticulata*), and kumquat (*F. hindsii*) accessions. The three 153-bp alleles clustered together and were unresolved. The bootstrap value for this group was highly supported with a value of 100% (Fig. 1c). Homoplasy also was observed in the 153-bp alleles. The repeat motifs were different for each 153-bp allele with repeat motifs of $(GT)_3ATCT(GT)_3(AT)_2$, $(GT)_3TTCT(GT)_3(AT)_2$, and $(GT)_3TTCT(GT)_4AT$. The 153-bp alleles had a few point mutations (C-to-G and A-to-C transversions in the flanking region, G-to-A transition in the microsatellite) that were specific to this allele size class and occurred in two separate genera and species (*Fortunella hindsii* and *C. reticulata*). In a population structure analysis (Barkley et al. 2006), CRC 2867 *C. reticulata* was found to be a hybrid having approximately 60% of its alleles derived from kumquats and only 40% from mandarins, which may explain why these accessions classified in separate genera share allele-specific mutations.

Fig. 1 **a** Strict consensus tree produced from the sequence data of alleles (155, 158, 161, and 164) from marker CCT01. **b** Strict consensus tree produced from the sequence data of alleles (103, 112, 115, and 118) from marker cAGG9. **c** Strict consensus tree produced from the sequence data of alleles (151, 153, 167, 171, and 173) from marker GT03. **d** Strict consensus tree produced from only the flanking sequences of microsatellite alleles derived from marker GT03. The names on the termini of all branches include allele size, CRC number, cultivar name, and taxonomic group, respectively ▶

The 151-bp allele clustered with the 153-bp alleles (Fig. 1c). This 151-bp allele derived from 'Japansche' (CRC 2875) has lost the TTCT interrupt contained within the microsatellite that all other taxa share and has more GT repeats than the 153-bp alleles. This 151-bp allele demonstrates an exception to the stepwise mutation model. One would expect that a dinucleotide repeat microsatellite locus evolving in a purely stepwise manner would contain a 2-bp deletion of a repeat motif when comparing a 153-bp allele to a 151-bp allele. However, it is also possible that the 151-bp allele is the ancestral allele and all the other alleles gained the TTCT interrupt contained within the microsatellite.

The allele sequences were edited to remove the microsatellite to examine the influence of the flanking sequence on the evolutionary relationships among taxa (Fig. 1d). Once again, the number of parsimony informative sites was drastically reduced, from 18 to 5, suggesting that the majority of the variation is contained within the repeat motif. Even though removing the microsatellite region reduced the number of parsimony informative sites, the resulting tree was fairly similar to the tree obtained with the entire sequence in which alleles of the same size class clustered together. All of the 173-bp alleles (all *P. trifoliata* taxa) and two of the 171-bp alleles (*P. trifoliata* taxa) clustered together and were undistinguishable due to a shared parsimonous site (C-to-T) in the flanking sequence, which would be expected based on taxonomy and phylogenies based on molecular marker data (Nicolosi et al. 2000; Barkley et al. 2006; Pang et al. 2007). Four of the five 167-bp alleles derived from the citrus relatives clustered together and were unresolved. The other 167-bp allele from a pummelo (*C. maxima*) accession clustered with the remaining 171-bp alleles produced from ancestral *Citrus* taxa. Since the overall clustering pattern was similar to the


a
164 1225 Hunnan pummelo
164 578 Flemings Shaddock pummelo
164 2341 Karn Lau Yau pummelo
164 2355 Kao Panne pummelo
164 3066 Sour pummelo
161 2341 Karn Lau Yau pummelo
161 3793 Unnamed papeda
161 3237 Unnamed kumquat
161 448 Moanalua pummelo
161 3055 Bengal citron
161 3816 Kinkoji Unshiu mandarin
161 644 Philippine pummelo
158 644 Philippine pummelo
158 661 Indian citron
158 3816 Kinkoji Unshiu mandarin
158 2554 Barnes trifoliate
155 3056 Unnamed papeda
63
158 3066 Sour pummelo
158 3055 Bengal citron
158 3237 Unnamed kumquat
53
155 3780 Unnamed citron
58
155 3793 Unnamed papeda
155 3797 Unnamed papeda
64
158 300 Parsons Special mandarin
158 2355 Kao Panne pummelo
b
118 2554 Barnes trifoliate
118 4008 Seedling trifoliate
118 2240 Siamese Acidless pummelo
118 3816 Kinkoji Unshiu mandarin
118 1208 Roedings Pink pummelo
115 3845 King mandarin
115 3055 Bengal citron
115 279 Clementine mandarin
115 1208 Roedings Pink pummelo
115 3816 Kinkoji Unshiu mandarin
63
115 3326 Scarlet Emperor mandarin
112 3147 Unnamed mandarin
112 3469 Hanayu papeda
112 3150 Unnamed mandarin
64
112 131 Santa Barbara rangpur
112 3845 King mandarin
103 3163 Indian Wild Orange mandarin
103 131 Santa Barbara rangpur
94
103 138 Indian citron
103 2875 Japansche citron
103 3527Hiawassiecitron

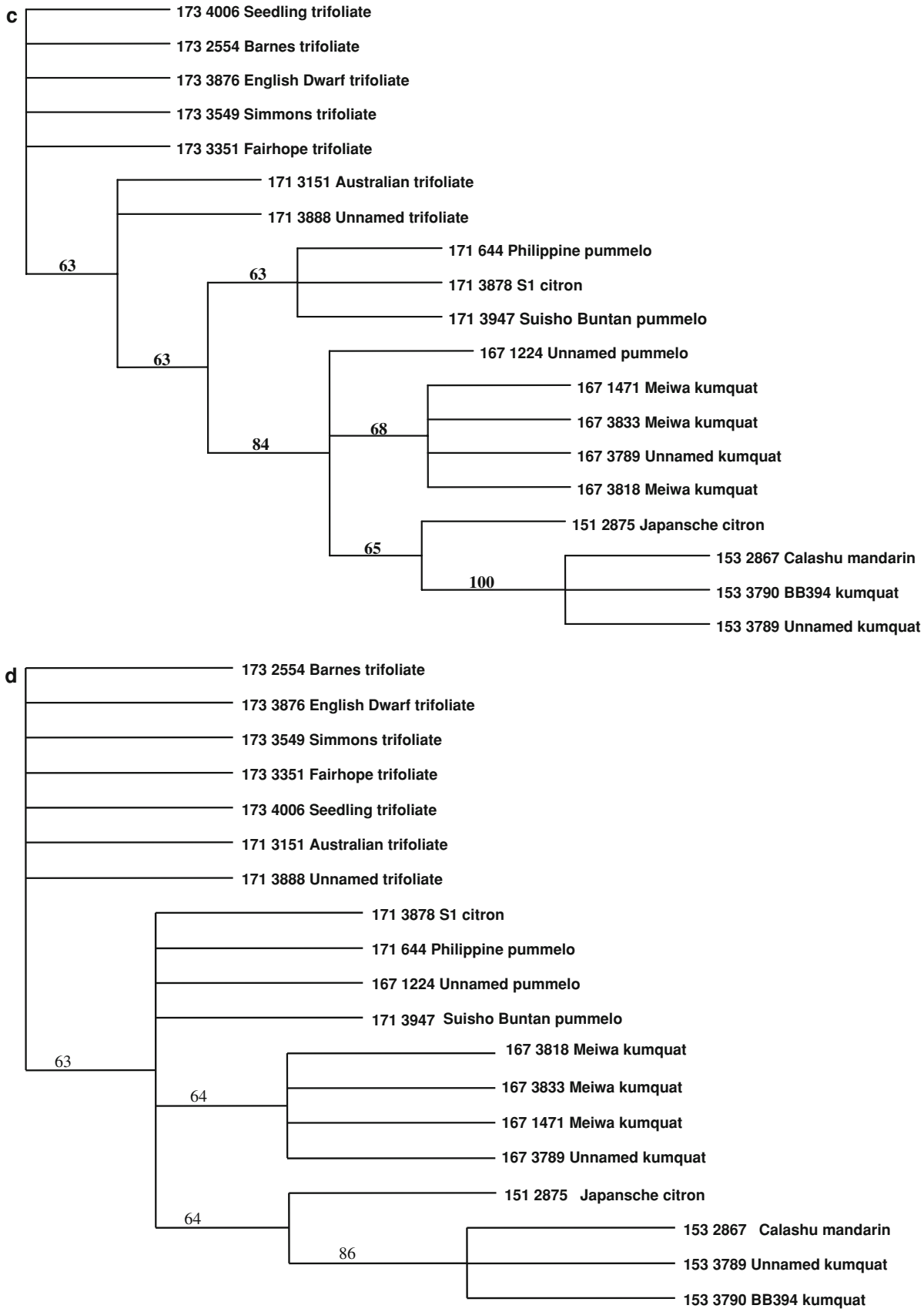


Fig. 1 continued

tree produced with the entire sequence, this would imply that the flanking sequence polymorphisms significantly contributed to the evolutionary relationships between these allele sequences. This further suggests that the flanking sequence of GT03 could be used effectively as a marker to deduce evolutionary relationships between taxa.

Networks/gene genealogies

TCS version 1.13 (Clement et al. 2000) was employed to produce a gene genealogy or network for each marker. This program reduces the data set by eliminating duplicate haplotypes and calculates the probability of parsimony for all pairwise combinations until the probability surpasses 0.95. The output from TCS was used to diagram a gene genealogy for each marker (Fig. 2a–c). The alleles used in this study were not randomly sampled, and thus, the ancestral allele chosen by TCS probably does not represent an actual ancestral allele. Locus CCT01 had the most complex network (Fig. 2a) with relatively few duplicate haplotypes. Only 4 alleles out of 25 had identical haplotypes at locus CCT01. The 158-bp allele produced from 'Indian' citron CRC 661 was picked as the haplotype from which the other haplotypes could be derived with the least amount of change. This analysis also suggests that the 158-bp alleles from CRC 644 pummelo may not have arisen in a stepwise manner, but instead gained and lost a repeat element from the 158-bp ancestral haplotype CRC 661 or possibly lost two repeat units from the 164-bp allele. Locus cAGG9 had a fairly simple network (Fig. 2b), probably due to the comparatively few point mutations observed. In contrast to marker CCT01, many of the alleles of the same size class had identical sequences, and thus, were reduced to a single haplotype such as the 118-bp allele. Even though locus GT03 had many cases of homoplasy and numerous parsimony informative sites (discussed previously), the network for GT03 was also fairly simple and several alleles of the same size class were reduced to a single haplotype (Fig. 2c).

Single-site polymorphism

The total number of single-site polymorphisms was calculated by examining the number of point mutations that occurred in each allele size class (Table 2). The 151-bp allele at locus GT03 was not included because only one 151-bp allele was cloned. Locus CCT01 had the most single-site polymorphisms with a total of 22 among the four allele size classes. On the other hand, locus cAGG9 had the fewest total single-site polymorphisms with seven observed for all four alleles. The 118-bp allele produced by primers targeting locus cAGG9 was the only allele in which no single-site polymorphisms were observed, indicating high sequence conservation. This allele also had a relatively low frequency of 0.038 in a population of 370 citrus accessions (Table 2), which may be why this sequence was completely conserved. The 158-bp allele at locus CCT01 had 10 single-site polymorphisms. This was the most observed in any allele size class and it also had the highest allele frequency (0.8469) of all the alleles targeted, suggesting that alleles that are common in the population may have a higher substitution rate or represent more ancient alleles. For loci CCT01 and cAGG9, the number of single-site mutations generally increased with allele frequency (Fig. 3). However, more alleles of different frequencies would need to be included to validate this trend. Single-site polymorphisms were also calculated for each locus examining all the sequence data collected per marker as opposed to evaluating each allele size class. The total number of single-site polymorphisms for all three loci was 30 with 16, 7, and 6 single-site polymorphisms observed at markers CCT01, GT03, and cAGG9, respectively.

Frequency of gaps and substitutions

The frequencies of gaps and substitutions were compared for all of the alleles at each of the three microsatellite loci examined to determine if gaps or base substitution mutations were more frequent. The alleles of marker cAGG9 had very few base substitutions (11) in comparison to the alleles produced from markers GT03 and CCT01, which had 40 and 34 base substitutions, respectively. Therefore, the frequency of base substitutions was 0.46, 0.83, and 1.26 per 100 bp for cAGG9, CCT01, and GT03. The mean frequency of base substitutions for all three loci was 0.85 per 100 bp. The number of trinucleotide gaps that were observed in the various sized alleles for markers cAGG9 and CCT01 were 41 and 37, respectively. Marker GT03 had a total of 61 dinucleotide gaps for all the alleles targeted at this locus. Therefore, given the total number of bases sequenced at marker cAGG9, the gaps occurred 3.73 times more frequently than the base substitutions. However, for markers GT03 and CCT01 the gaps occurred only 1.53 and 1.1 times more than base substitutions, respectively. Therefore, it appears that insertion or deletion of repeat elements occurs more frequently than base substitutions for the alleles at these three microsatellite loci assuming that changes in allele size are due to the addition or deletion of one repeat element at a time.

Sequence variation

DNA polymorphism from nucleotide sequence data generated from the three microsatellite markers was examined by using the program DnaSP version 3.5 (Rozas and Rozas 1999). The average number of different nucleotides per site

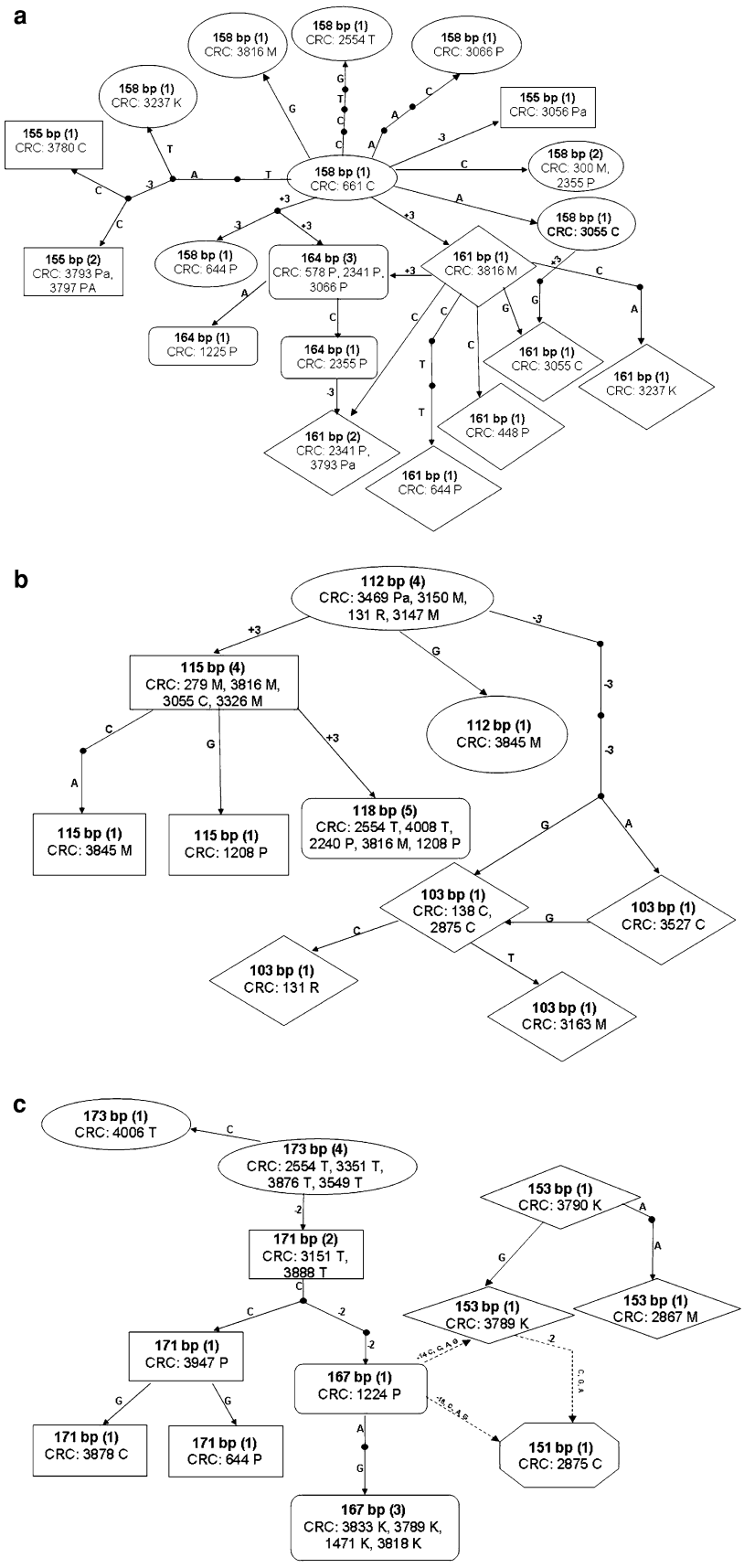


Fig. 2 **a** CCT01 network/gene genealogy constructed from TCS 1.13 (Clement et al. 2000). **b** cAGG9 network/gene genealogy. **c** GT03 network/gene genealogy. The *branches* are labeled to denote the changes between alleles, and the *arrows* denote the direction of the change shown on the diagram. Each allele size is denoted by a different shape (e.g., −158 bp = oval). The *text within the shapes* denotes the allele size, CRC number, and a letter (*C* citron, *K* kumquat, *M* mandarin, *P* pummelo, *Pa* papeda, *R* rangpur, and *T* trifoliate) to denote the group in which the particular CRC number is classified. The order of mutations denoted on branches is arbitrary

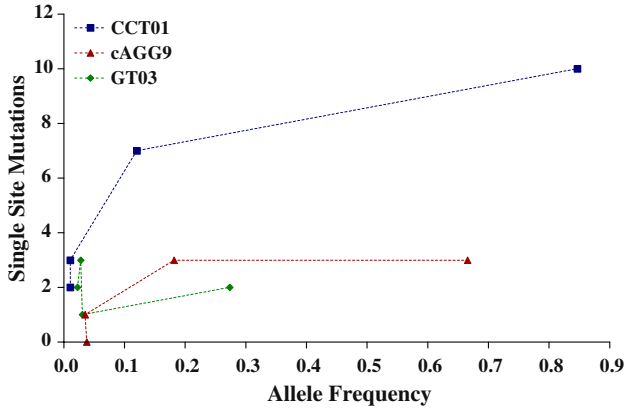


Fig. 3 Scatter plot showing the general relationship of allele frequency (*x*-axis) compared to the number of single-site mutations (*y*-axis) for each of the three markers used in this study

between sequences (π) ranged from 0.00730 for marker cAGG9 to 0.01624 for marker CCT01 with a mean value of 0.01324. The nucleotide diversity (π) for marker GT03 was 0.01618. The average number of nucleotide differences (K) for cAGG9, GT03, and CCT01 was 0.781, 2.281, and 2.533, respectively. Theta (θ) was calculated per site from the number of polymorphic sites and also calculated per site from the total number of mutations. The values of θ calculated per site from the number of segregating sites for markers cAGG9, GT03, and CCT01 were 0.01299, 0.02029, and 0.03565, respectively. The values of θ per site estimated from the total number of mutations were 0.01559, 0.02029, and 0.03565 for cAGG9, GT03, and CCT01, respectively. In general, these values calculated for sequence variation can only be considered estimates due to the selection of samples based on allele frequency and maximizing ancestral taxa, which included multiple genera, species, and a few hybrids. Further work could include analyzing multiple individuals from a single taxon to obtain precise measures of sequence variation.

Discussion

Microsatellite markers are frequently used in molecular genetic studies because they are codominant, polymorphic, and ubiquitous in eukaryotic genomes. These markers have been extensively used for assessing genetic diversity, fingerprinting, determining parentage, forensics, construction of genetic linkage maps, and phylogenetic analysis. SSR markers can be effective tools for most all of these research objectives; however, due to extensive homoplasy, they may fail or lead to incorrect conclusions in phylogenetic analysis when evaluating divergent intraspecific, interspecific, or intergeneric relationships. In principle, homoplasy is tightly linked to the mechanisms that cause mutations that produce new alleles, and hence, homoplasy is also coupled to the underlying mutation model (Lia et al. 2007). The amount of homoplasy in microsatellite alleles is generally thought to increase with increasing time of divergence among taxa (van Oppen et al. 2000) and increasing allele size since longer repeats are less stable than shorter repeats (Anmarkrud et al. 2008). Previous interspecific studies of the sequence content of microsatellite alleles have revealed prevalent homoplasy including loss of the targeted repeat motif; hence, it has been suggested that these markers may not be useful for phylogenetic analysis above the species level (Ochieng et al. 2007; Tesfaye et al. 2007). However, since genetic distance measures (used for phylogenetic construction) are averaged over several loci, a few cases of homoplasy would probably not invalidate the average relationships between taxa, but the effect or amount of allowable homoplasy in microsatellite alleles for phylogenetic construction is currently unknown.

Genetic distance measures used to construct a phylogeny assume that allelic size class is an indication of phylogenetic affinity (Orti et al. 1997). Furthermore, phylogenetic reconstruction is based on the assumption that mutations between individuals increase as the time increases since they diverged from a common ancestor (Holder and Lewis 2003). Therefore, if microsatellite alleles arise by convergent or parallel evolution in which different lineages acquire the same trait, revert back to their ancestral states, or do not contain similar sequence content for alleles that are identical in state, the ability to infer patterns of evolutionary history can be affected (Adams et al. 2004). Measuring homoplasy caused by convergent evolution can be difficult without evaluating mutations in known pedigrees. (Currently, extensive pedigree information in citrus is limited). On the other hand, homoplastic alleles that are identical in state (IIS), but not identical by descent (IBD), and thus, contain different sequences for the same sized alleles can be easily evaluated by determining the sequence content. Since allelic homoplasy and hidden motifs within alleles have been demonstrated in several microsatellite studies (Grimaldi and Crouau-Roy 1997; Primmer and Ellegren 1998; Viard et al. 1998; Culver et al. 2001; Hale et al. 2004), one needs to be careful in interpreting results from microsatellite data based solely on allele size data particularly for distantly related taxa.

Because citrus taxonomy can be somewhat debatable and microsatellite markers are suggested to be employed only for intraspecific relationships, our goal in this study was to evaluate what changes might exist in these alleles over what is assumed to be divergent citrus taxa (based on current taxonomy and previous molecular marker studies). The main obstacles in classifying citrus are disagreement on whether hybrids among naturally occurring forms should be assigned species rank (Roose et al. 1995),

repeated cross pollination among taxa, and nucellar embryony, which perpetuates hybrid taxa (Scora, 1975). The results from sequencing microsatellite alleles in *Citrus* spp. and its two closest relatives showed that the expected repeat motifs were present, albeit sometimes slightly modified, even when evaluating alleles derived from separate taxonomic genera and species. Since most *Citrus* taxa can freely cross with one another, conservation of microsatellites among interspecific accessions is not too unexpected, whereas studies of other species and genera have not always observed microsatellite repeat preservation when examining distant taxonomic relatives (Chen et al. 2002). Moreover, the microsatellite alleles generally, but not always, provided information about their relatedness in that same sized alleles clustered together; although they did not always display identity between alleles of the same size class. This suggests that employing microsatellite markers in the genus *Citrus*, *Poncirus*, and *Fortunella* may generate valid phylogenetic inferences when calculating genetic distances using mutation models that assume some homoplasy may occur. [Currently, several mutation models used for the analysis of microsatellite markers to calculate genetic distance such as the stepwise mutation model, K-allele model (Kimura 1968), and the two-phase model (Di Rienzo et al. 1994) assume that some homoplasy may occur; however, the infinite allele model does not take into account that homoplasy may occur (Estoup et al. 2002)]. Additionally, since preservation of microsatellite motifs among distant species or genera is not always typical (Chen et al. 2002), this may suggest that as hypothesized by Mabberley (1997) and Bayer et al. (2009), species and genera rank may be over-inflated due to the commercial value of citrus. However, because homoplasy is thought to increase with time of divergence, it is also possible that the divergence time between species and genera has not been extensive enough to allow mutations to accumulate, and thus, significantly alter the sequence of these microsatellite alleles. Another possible explanation could be that these microsatellites are somewhat stable in distant citrus taxa due to their small size since larger repeat motifs have been shown to have more hidden sequence motifs than shorter alleles (Anmarkrud et al. 2008). Consequently, one may be able to control or reduce the amount of homoplasy in microsatellite alleles by intentionally selecting microsatellites with shorter repeat elements.

In conclusion, this work demonstrates that variation among microsatellite alleles in the genus *Citrus* and two related genera (*Poncirus* and *Fortunella*) were fairly consistent with the stepwise mutation model. Interallelic variation in all of the targeted alleles at these three loci with one notable exception could all be explained by an expansion or contraction of repeat units. No indels were detected in the flanking sequence as seen in several previous studies (Orti et al. 1997; Makova et al. 2000; Matsuoka et al. 2002). This work suggests that microsatellites can be a useful tool for evaluating *Citrus* species and two related genera since repeat motifs were reasonably well retained. Homoplasy was detected at all three loci but was most prevalent in markers GT03 and CCT01; consequently, the number of microsatellite alleles is clearly an underestimate of the number of sequence variants present. Therefore, this suggests that allele size data do not always represent the true level of genetic diversity present in *Citrus* and two related genera. In general, as the allele frequency increased in the population so did the number of single-site mutations, which in turn generated some of the observed homoplasy; however, more work needs to be done with a range of alleles at different frequencies in the population and the inclusion of more markers to validate this trend. In addition, sequencing these alleles demonstrated new genetic variation, some of which was specific to certain genera or species that would not have been revealed based on size alone, which with further testing could be used to develop SNP markers to distinguish individual accessions or a particular species. Overall, this study along with others adds to the growing body of evidence that microsatellite alleles that are similar in size are not necessarily characterized by identical sequence content or do not necessarily contain the expected microsatellite repeats. Thus, careful examination of the sequence content of alleles should be performed prior to making any conclusions about the assumed evolutionary relationships between accessions.

Acknowledgements We would like to thank the US Department of Agriculture for the funding to support this project. In addition, we thank colleagues at USDA-ARS Plant Genetic Resource Conservation Unit in Griffin, GA, for reviewing this manuscript and providing suggestions.

References

Adams RI, Brown KM, Hamilton MB (2004) The impact of microsatellite electromorph size homoplasy on multilocus population structure estimates in a tropical tree (*Corythophora alta*) and an anadromous fish (*Morone saxatilis*). Mol Ecol 13:2579–2588

Angers B, Estoup A, Jarne P (2000) Microsatellite size homoplasy, SSCP, and population structure: a case study in the freshwater snail *Bulinus truncatas*. Mol Biol Evol 17:1926–1932

Anmarkrud JA, Kleven O, Bachmann L, Lifjeld JT (2008) Microsatellite evolution: mutations, sequence variation, and homoplasy in the hypervariable avian microsatellite locus HrU10. BMC Evol Biol 8:138

Barkley NA, Roose ML, Krueger RR, Federici CT (2006) Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). Theor Appl Genet 112:1519–1531

Barrett HC, Rhodes AM (1976) A numerical taxonomic study of affinity relationships in cultivated *Citrus* and its close relatives. Syst Bot 1:105–136

Bayer RJ, Mabberley DJ, Morton C, Miller CH, Sharma IK, Pfeil BE, Rich S, Hitchcock R, Sykes S (2009) A molecular phylogeny of the orange subfamily (Rutaceae: Aurantioideae) using nine cpDNA sequences. Am J Bot 96:668–685

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 28:1040–1050

Chen X, Cho YG, McCouch SR (2002) Sequence divergence of rice microsatellites in *Oryza* and other plant species. Mol Genet Genomics 268:331–343

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. Mol Ecol 9:1657–1659

Culver M, Menotti-Raymond MA, O'Brien SJ (2001) Patterns of size homoplasy at 10 microsatellite loci in pumas (*Puma concolor*). Mol Bio Evol 18:1151–1156

Curtu AL, Finkeldey R, Gailing O (2004) Comparative sequencing of a microsatellite locus reveals size homoplasy within and between European oak species (*Quercus* spp.). Plant Mol Bio Reporter 22:339–346

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci USA 91:3166–3170

Estoup A, Cornuet JM (1999) Microsatellite evolution: inferences from population data. In: Goldstein DB, Schlötterer C (eds) Microsatellites evolution and applications. Oxford University Press, New York, pp 49–65

Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetic analysis. Mol Ecol 11:1591–1604

Fang DQ, Roose ML, Krueger RR, Federici CT (1997) Fingerprinting trifoliate orange germplasm accessions with isozymes, RFLPs, and inter-simple sequence repeat markers. Theor Appl Genet 95:211–219

Federici CT, Fang DQ, Scora RW, Roose ML (1998) Phylogenetic relationships within the genus *Citrus* (Rutaceae) and related genera as revealed by RFLP and RAPD analysis. Theor Appl Genet 94:812–822

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463–471

Grimaldi MC, Crouau-Roy B (1997) Microsatellite allelic homoplasy due to variable flanking sequences. J Mol Evol 44:336–340

Hale ML, Borland AM, Gustafsson MHG, Wolff K (2004) Causes of size homoplasy among chloroplast microsatellites in closely related *Clusia* species. J Mol Evol 58:182–190

Holder M, Lewis PW (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev (Genet) 4:275–284

Xie H, Sui Y, Chang FU, Xu FY, Ma RC (2006) SSR allelic variation in almond (*Prunus dulcis* Mill.). Theor Appl Genet 112:366–372

Kimura M (1968) Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet Res 11:247–269

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Bio Evol 4:203–221

Lia VV, Bracco M, Gottlieb AM, Poggio L, Confalonieri VA (2007) Complex mutational patterns and size homoplasy at maize microsatellite loci. Theor Appl Genet 115:981–991

Mabberley DJ (1997) A classification for edible *Citrus* (Rutaceae). Telopea 7:167–172

Makova K, Nekrutenka A, Baker RJ (2000) Evolution of microsatellite alleles in four species of mice (Genus *Apodemus*). J Mol Evol 51:166–172

Matsuoka Y, Mitchell SE, Kresovich S, Goodman M, Doebley J (2002) Microsatellites in *Zea*—variability, patterns of mutations and use for evolutionary studies. Theor Appl Genet 104:436–450

Nicolosi E, Deng ZN, Gentile A, La Malfa S, Continella G, Tribulato E (2000) Citrus phylogeny and genetic origin of important species as investigated by molecular markers. Theor Appl Genet 100:1155–1166

Ochieng JW, Muigai AW, Ude GN (2007) Phylogenetics in plant biotechnology: principles, obstacles and opportunities for the resource poor. Afr J Biotech 6:639–649

Orti G, Pearse DE, Avise JC (1997) Phylogenetic assessment of length variation at a microsatellite locus. Proc Natl Acad Sci USA 94:10745–10749

Pang XM, Hu CG, Deng XX (2007) Phylogenetic relationships within *Citrus* and its related genera as inferred from AFLP markers. Genet Res Crop Evol 54:429–436

Peakall R, Gilmore S, Keys W, Morgante M, Rafalski A (1998) Cross-species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Mol Biol Evol 15:1275–1287

Primmer CR, Ellegren H (1998) Patterns of molecular evolution in avian microsatellites. Mol Bio Evol 15:997–1008

Roose ML, Soost RK, Cameron JW (1995) Citrus. In: Smart J, Simmonds NW (eds) The evolution of crop plants, 2nd ed. Longman, Essex, pp 443–449

Rossetto M, McNally J, Henry RJ (2002) Evaluating the potential of SSR flanking regions for examining taxonomic relationships in the Vitaceae. Theor Appl Genet 104:61–66

Rousset F (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 142:1357–1362

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Schlötterer C (1998) Genome evolution: are microsatellites really simple sequences? Curr Biol 8:R132–134

Scora RW (1975) On the history and origin of *Citrus*. Bull Torr Bot Club 102:369–375

Shriver MD, Jin L, Boerwinkle E, Deka R, Ferrell RE, Chakraborty R (1995) A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol Biol Evol 12:914–920

Swingle WT, Reece PC (1967) The botany of *Citrus* and its wild relatives. In: Reuther W, Webber HJ, Batchelor LD (eds) The citrus industry, vol 1. University of California Press, Berkeley, pp 190–430

Swofford DL (2003) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland

Symonds VV, Lloyd AM (2003) An analysis of microsatellite loci in *Arabidopsis thaliana*: mutational dynamics and application. Genetics 165:1475–1488

Tanaka T (1977) Fundamental discussion of *Citrus* classification. Stud Citrol 14:1–6

Taylor JS, Sanny JS, Breden F (1999) Microsatellite allele size homoplasy in the guppy (*Poecilia reticulata*). J Mol Evol 48:245–247

Tesfaye K, Borsch T, Govers K, Bekele E (2007) Characterization of *Coffea* chloroplast microsatellites and evidence for the recent divergence of *C. Arabica* and *C. eugeniodes* chloroplast genomes. Genome 50:1112–1129

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Valdes AM, Slatkin M, Freimer NB (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737–749

van Oppen MJH, Rico C, Turner GF, Hewitt GM (2000) Extensive homoplasy, nonstepwise mutations and shared ancestral polymorphism at a complex microsatellite locus in Lake Malawi cichlids. Mol Biol Evol 17:489–498

Viard F, Franck P, Dubois MP, Estoup A, Jarne P (1998) Variation of microsatellite size homoplasy across electromorphs, loci and populations in three invertebrate species. J Mol Evol 47:42–51

Zhu Y, Queller D, Strassman J (2000) A phylogenetic perspective on sequence evolution in microsatellite loci. J Mol Evol 50:324–338